

Отчёт о выполнении работы

Выполнил: Солдатенков Антон

Цель работы:

Получить навыки организации взаимодействия различных моделей машинного обучения и реализации единого жизненного цикла эксплуатации (pipeline).

Задачи работы:

1. Использовать модель кластеризации, реализованную в лабораторной работе №3, для генерации размеченного набора данных, полученного также в лаб. раб. №3, по кластерам (классам).
2. Провести анализ сгенерированных данных, разделить на обучающую и тестовую выборки.
3. Реализовать модель классификации сгенерированных данных с любым изученным алгоритмом машинного обучения.
4. Реализовать модель линейной регрессии для каждого вектора признаков, соответствующему метке класса, полученной на выходе модели классификации.
5. Вывести метрики трёх моделей, то есть построить матрицы ошибок, графики регрессии и т.д.
6. Использовать API Apache Spark/PySpark.

Ход работы

1. Згружены полученные при выполнении работы №3 результаты кластеризации
2. Выполнен разведочный анализ данных (EDA), данные разбиты на обучающую и тестовую выборки
3. Средствами SparkML реализованы указанные в задачах работы модели классификации и регрессии
4. Выведены метрики полученных моделей на обучающей и тестовой выборках (файл `log.txt`), построены графики обучения (сохранены в папку `plots`)